Representation of Word Meaning in the Intermediate Projection Layer of a Neural Language Model

Steven Derby[1] **Paul Miller**[1] **Brian Murphy**[1,2] **Barry Devereux**[1]

[1] Queen's University Belfast, Belfast, United Kingdom
[2] BrainWaveBank Ltd., Belfast, United Kingdom
{sderby02, p.miller, brian.murphy, b.devereux}@qub.ac.uk

Abstract

Performance in language modelling has been significantly improved by training recurrent neural networks on large corpora. This progress has come at the cost of interpretability and an understanding of how these architectures function, making principled development of better language models more difficult. We look inside a state-of-the-art neural language model to analyse how this model represents high-level lexico-semantic information. In particular, we investigate how the model represents words by extracting activation patterns where they occur in the text, and compare these representations directly to human semantic knowledge.

1 Introduction & Related Work

Language modelling involves learning to predict the next word in a sequence of words, using large text corpora as the training input. Language models must therefore learn to represent information from the preceding context which is relevant for future word prediction, and, intuitively, this should include information about the syntactic structure of the context and the meanings of constituent words. Today's state-of-the-art language models make use of Recurrent Neural Networks (RNNs) with Long Short-Term Memory cells (LSTMs) (Hochreiter and Schmidhuber, 1997) which can handle time series information by remembering salient information over latent variables (Mikolov et al., 2010). Because of their wide applicability, there has been much interest in developing a better understanding of the inner workings of RNN models, and, in particular, researchers have investigated how syntactic knowledge is encoded and processed by such networks (Dyer et al., 2016; Linzen et al., 2016; Jozefowicz et al., 2016; McCoy et al., 2018; Gulordava et al., 2018). Karpathy et al. (2015) performed an in-depth analysis of the types of errors RNN's make, in order to understand how recurrent mechanisms can encode long-term dependency information. Linzen et al. (2016) present a more direct analysis by examining LSTM language models' ability to understand difficult long-range dependencies such as the form of a verb linked to a noun subject. Recently, researchers have started to study the semantic embeddings generated by these networks (Chrupała et al., 2015), especially for those focused on encoding visual grounding (Kiela et al., 2017; Yoo et al., 2017). However, compared to syntax, there has been relatively less work on how LSTM networks represent lexical semantic knowledge.

In this work, we evaluate latent semantic knowledge present in the LSTM activation patterns produced before and after the word of interest. We evaluate whether these activations predict human similarity ratings, human-derived property knowledge, and brain imaging data. In this way, we test the model's ability to encode important semantic information relevant to word prediction, and it's relationship with human cognitive semantic representations.

2 Language Model Data

We make use of a state-of-the-art LSTM neural language model known as lm_1b (Jozefowicz et al., 2016), which consists of two LSTM layers followed by low-dimensional projections. To construct representations from the language model's LSTM projection layer, we first select a subset of 62.5 million sentences from the One Billion Word dataset (Chelba et al., 2013). We then choose a predefined set of target words, based on the overlap of words in the lm_1b vocabulary with words used in three evaluation datasets, described in Section 3. To derive a model of the lexical representation for each of our target words using the language model, we sample 100 sentences for each word in which that word occurs, and process each of those sentences using lm_1b. More specifically, at the location in the sentence where the specific

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 362–364
Brussels, Belgium, November 1, 2018.

word of interest has just been processed, we record the 1024-dimensional projection of the activations of the first LSTM layer in the network and then average all these vectors (from 100 sentences) to get the final vector. On the assumption that the effects of context "average out" over the 100 different sampled sentences for each word, we take the average vector to be a representation of the lexical content of the concept, independent of context. Furthermore, we also build a model of lexical representation by recording the LSTM activations at the word just *before* the target word is presented to the network.

3 Experiments & Results

3.1 Comparison to Similarity Judgments

We first investigate how well similarities between our model vectors predict human similarity judgments. We use WordSim353 (Finkelstein et al., 2001) a set of 353 pairs of words along with human ratings. We split WordSim353 into semantic similarity and semantic relatedness datasets, following Agirre et al. (2009). On the hypothesis that the representations we derived from the language model reflect lexical content, we predicted that similarity, as calculated from the model, would more closely correspond to semantic similarity (i.e. shared hypernyms) than semantic relatedness. We also anticipated that correlations with human judgments would be stronger for the 'after' model than the 'before' model, since the word explicitly affects activations in the network only after it is encountered (however, the 'before' model provides an interesting test of whether lexical information can be predicted, drawing an analogy with models of human language comprehension (Kuperberg, 2016)).

For both the before and after models, correlations were stronger for the human semantic similarity ratings than for semantic relatedness, with the strongest correlation achieved for the 'after' model and similarity ratings (r=0.30). Furthermore, the after model more closely corresponded to the human similarities than the before model, though the before model still shows some correlation (r=0.21), indicating that the model may indeed encode information about upcoming concepts before they occur.

3.2 Property Knowledge Prediction

To directly investigate how the language model encodes lexico-semantic content, we analysed whether the derived lexical representations can predict human-derived properties of the same concepts. We used a dataset of human-elicited property knowledge (the CSLB norms; Devereux et al. (2014)), which lists semantic properties for concepts (e.g. *leaf* has the properties *is-green* & *grows-on-trees*). To test how well the model representations can predict these properties, we largely follow Collell and Moens (2016) and Lucy and Gauthier (2017). For each property, we train an $L2$-regularized logistic regression to predict whether that property is true for a given concept. We train two sets of logistic regression models to predict properties from the vectors in the 'before' and 'after' models. We use 5-fold cross validation with stratified sampling to ensure at least one positive case occurs in the validation fold. To get the final score of the decodability of a property for each model, we average the F1 scores over each test fold. Interestingly, semantic features were more decodable before the noun than afterwards.

3.3 Comparison to Brain Imaging Data

We compared the before and after representations from the language model to fMRI and MEG brain imaging data for 60 concepts available in BrainBench (Xu et al., 2016). We use the "2 vs 2" test described in Xu et al. (2016) for all pairs of concepts to measure the correspondence between the models and the brain data. The 'before' and 'after' models perform similarly, though (somewhat surprisingly) the before model performs slightly better on fMRI data than the after model. However, both models perform above chance, indicating that these models are correlated with brain representations of the same nouns.

4 Conclusions

Our results suggest that LSTM language models not only encode probabilistic syntactic knowledge but also represent the semantic content of words in a way which is at least somewhat consistent with measures of human conceptual knowledge. Language models' ability to predict human property knowledge allows us to draw initial comparisons between these models and activation (and pre-activation) of lexical information in human language comprehension.

Acknowledgements

This work was partly funded by a Microsoft Azure for Research Award.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Grzegorz Chrupała, Akos Kádár, and Afra Alishahi. 2015. Learning language through pictures. *arXiv preprint arXiv:1506.03694*.

Guillem Collell and Marie-Francine Moens. 2016. Is an image worth more than a thousand words? on the fine-grain semantic differences between visual and linguistic representations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2807–2817. The COLING 2016 Organizing Committee.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46(4):1119–1127.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. 2016. Recurrent neural network grammars. *arXiv preprint arXiv:1602.07776*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2017. Learning visually grounded sentence representations. *arXiv preprint arXiv:1707.06320*.

& Jaeger T. F. Kuperberg, G. R. 2016. What do we mean by prediction in language comprehension? *anguage, Cognition and Neuroscience*, 31(1):32–59.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *arXiv preprint arXiv:1611.01368*.

Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. *arXiv preprint arXiv:1705.11168*.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society.*

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Haoyan Xu, Brian Murphy, and Alona Fyshe. 2016. Brainbench: A brain-image test suite for distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2017–2021.

Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. 2017. Improving visually grounded sentence representations with self-attention. *arXiv preprint arXiv:1712.00609*.